ROBUST GLOBAL MOTION ESTIMATION WITH MATRIX COMPLETION

Federica Arrigoni[a]*, Beatrice Rossi[b], Francesco Malapelle[c], Pasqualina Fragneto[b], Andrea Fusiello[c]

[a] Dipartimento di Matematica, Università degli Studi di Milano, Milano, Italy - federica.arrigoni@studenti.unimi.it
[b] ASTLab, STMicroelectronics, Agrate Brianza (MB), Italy - (name.surname)@st.com
[c] DIEGM, Università degli Studi di Udine, Udine, Italy - (name.surname)@uniud.it

Commission V

KEY WORDS: Camera orientation, Structure from motion, Epipolar geometry, Block adjustment

ABSTRACT:

In this paper we address the problem of estimating the attitudes and positions of a set of cameras in an external coordinate system. Starting from a conventional global structure-from-motion pipeline, we present some substantial advances. In order to detect outlier relative rotations extracted from pairs of views, we improve a state-of-the-art algorithm based on cycle consistency, by introducing cycle bases. We estimate the angular attitudes of the cameras by proposing a novel gradient descent algorithm based on low-rank matrix completion, that naturally copes with the case of missing data. As for position recovery, we analyze an existing technique from a theoretical point of view, providing some insights on the conditions that guarantee solvability. We provide experimental results on both synthetic and real image sequences for which ground truth calibration is provided.

1 INTRODUCTION

Block adjustment has a pivotal role in modern Photogrammetry. The same technique is referred to as *Structure from Motion* (SfM) in Computer Vision: given multiple images of a stationary scene, the goal is to recover both *scene structure*, i.e. 3D coordinates of object points, and *camera motion*, i.e. the exterior orientation (position and attitude) of the photographs. It is assumed that the interior parameters of the cameras are known, namely the focal length and the coordinates of the principal point.

Structure-from-motion methods can be classified as: structure-first, like independent models block adjustments (e.g. (Crosilla and Beinat, 2002)), where first stereo-models are built and co-registered, structure-*and*-motion methods, such as bundle block adjustment (e.g. (Triggs et al., 2000)), resection-intersection methods (Brown and Lowe, 2005, Snavely et al., 2006), hierarchical methods (Gherardi et al., 2010, Ni and Dellaert, 2012)), where "structure" and "motion" are solved simultaneously, and – more recently – motion-first methods (Govindu, 2001, Martinec and Pajdla, 2007, Kahl and Hartley, 2008, Enqvist et al., 2011, Arie-Nachimson et al., 2012, Moulon et al., 2013) that first solve for the "motion" and then recover the "structure". All these motion-first methods are *global*, for they take into account the whole *epipolar graph*, whose nodes represent the views and edges link views having consistent matching points.

These global methods are usually faster than the others, while ensuring a fair distribution of the errors among the cameras, being global. Although the accuracy is worse than those achieved by bundle adjustment, these global methods can be seen as an effective and efficient way of computing approximate orientations to be subsequently refined by bundle adjustment.

In this paper we present a *robust* global structure-from-motion system, focusing in particular on the orientation process. First, the pairwise rotations extracted from the essential matrices are pruned by detecting and removing outliers, which arise from wrong two-view geometries caused (e.g.) by repetitive structures in the scene. To this end, we improve an algorithm based on the notion of cycle consistency (Enqvist et al., 2011) by introducing cycle bases.

*Corresponding author

In order to estimate the angular attitudes of the cameras, we formulate a gradient descent algorithm based on low-rank matrix completion, that naturally copes with the case of missing data, very frequent in practice, in which the epipolar graph is hardly complete.

As for position recovery, we analyze an existing technique (Arie-Nachimson et al., 2012) from a theoretical point of view, providing some insights on the conditions that guarantee solvability.

2 BACKGROUND

In this section we provide a brief summary of the main concepts in multi-view geometry, useful to define our method. A complete treatment of this subject can be found in (Hartley and Zisserman, 2004).

The camera model is the *pinhole* camera, which is described by its *centre* $\mathbf{c}$ and the *image plane*. The distance of the image plane from $\mathbf{c}$ is the *focal length*. The line from the camera centre perpendicular to the image plane is called the *principal axis*, and the point where the principal axis meets the image plane is called the *principal point*. A $3-$space point is projected onto the image plane through the line containing the point and the optical centre. Formally, the relationship between the homogeneous 3D coordinates $\mathbf{X}$ of a scene point and the homogeneous coordinates $\mathbf{x}$ of its projection onto the image plane is a mapping between projective spaces, called *central projection*:

$$P : \mathbb{P}^3 \to \mathbb{P}^2 \quad \mathbf{x} = P\mathbf{X}. \tag{1}$$

The mapping P appearing in (1) is called the *camera projection matrix* and can be expressed as $P = K[R|\mathbf{t}]$, where K is called the *calibration matrix*, that contains the *interior parameters*, and R, t are called the *exterior parameters*. The rotation matrix $R \in SO(3)$ and the translation vector $\mathbf{t} \in \mathbb{R}^3$ respectively describe the position and attitude of the camera with respect to an *external* (or global or control) coordinate system. The translation vector $\mathbf{t}$ is linked to the camera centre $\mathbf{c}$ through the formula $\mathbf{t} = -R\mathbf{c}$. The calibration matrix K encodes the transformation in the image plane from the so-called *normalized camera coordinates* to *pixel*

coordinates. It can be expressed as

$$K = \begin{bmatrix} rf & \gamma f & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

where f represents the focal length of the camera in terms of pixel dimensions (in the y direction), r is the aspect ratio, γ is the skew and $(p_x, p_y)^T$ are the coordinates of the principal point expressed in pixels.

The geometry of two images is the relative geometry of two different perspective views of the same 3D scene. It is usually referred to as *epipolar geometry*. Suppose a point $\mathbf{X}$ in $3-$space is imaged in two views, at $\mathbf{x}$ in the first, and $\mathbf{x}'$ in the second; $\mathbf{x}$ and $\mathbf{x}'$ are called *corresponding* or *matching points*. The camera centres, the 3-space point $\mathbf{X}$ and its images $\mathbf{x}$ and $\mathbf{x}'$ lie on a common plane, that is called the *epipolar plane*. As the position of $\mathbf{X}$ varies, the epipolar planes "rotate" about the *baseline*, that is the line connecting the camera centres. The image point $\mathbf{x}$ back-projects to a ray in 3-space (defined by the camera centre and $\mathbf{x}$) that is imaged as a line l' in the second view, called the *epipolar line*. The mapping $F : \mathbf{x} \mapsto l' = F\mathbf{x}$ associating a point in one image to its corresponding epipolar line in the other image is called the *fundamental matrix*. The corresponding point $\mathbf{x}'$ must lie on l', i.e. $\mathbf{x}'^T F\mathbf{x} = 0$. This relation is known as the *epipolar constraint*. If the calibration matrices K, K' of the cameras are known, then the *essential matrix* can be defined from the fundamental matrix as $E = K'^T F K$. The importance of the essential matrix is that it encodes the *relative motion* between the two cameras. In other words, the essential matrix admits the following decomposition

$$E = [\mathbf{t}]_\times R \quad (3)$$

where R and $\mathbf{t}$ respectively denote the relative rotation and translation between the cameras, and $[\mathbf{t}]_\times$ denotes the skew-symmetric matrix corresponding to the cross product with $\mathbf{t}$. Translation $\mathbf{t}$ can be recovered from E only up to an unknown scale factor which is inherited by the reconstruction.

If the scene is captured by $n \geq 2$ cameras, then for each available pair (i, j) we can compute the relative orientations $(R_{ij}, \mathbf{t}_{ij})$ starting from the essential matrix E_{ij}, according to (3). The problem now is to recover the exterior orientations, i.e., the rotation matrices $R_i \in SO(3)$ and the translation vectors $\mathbf{t}_i \in \mathbb{R}^3$ of the cameras such that the projection matrix of the $i-$th camera is

$$P_i = K_i[R_i|\mathbf{t}_i] \quad (4)$$

where K_i are the known calibration matrices.

3 OVERVIEW

Given n input images, we follow a standard global SfM pipeline.

1. A collection of key-points across the images is extracted and matched (typically by using SIFT (Lowe, 2004));
2. The essential matrices E_{ij} are computed from the matching points by using the $8-$point Algorithm in combination with RANSAC, and subsequently refined by using a Gauss-Newton type algorithm on the Essential Manifold, as explained in (Helmke et al., 2007). Finally, they are factorized via Singular Value Decomposition (SVD) to obtain the relative rotations R_{ij} of the cameras (Hartley and Zisserman, 2004);
3. Exterior camera orientation is computed as a sequence of two global optimizations. First, the attitude R_i of each camera is estimated, and then the positions $\mathbf{c}_i$ are recovered. This step is preceded by an outlier removal phase, in which wrong relative orientations are detected.
4. Optionally, the 3D coordinates of the key-points are computed by triangulation, and the quality of structure and motion estimation may be refined through bundle block adjustment.

Our contributions, which concern Step 3 of the above pipeline, are the following. First, we introduce a novel optimization method to estimate the exterior attitudes R_i of the cameras starting from the relative attitudes, that is described in Section 4. Secondly, we formulate a novel algorithm to remove the outliers among the relative rotations, presented in Section 5. Finally, in Section 6 we provide a theoretical analysis of the linear algorithm introduced in (Arie-Nachimson et al., 2012), that solve for exterior positions. The discussion carried out in the paper is supported by experimental results on both synthetic and real images, shown in Section 7. The conclusions along with possible further developments are presented in Section 8.

4 ATTITUDE ESTIMATION: A MATRIX COMPLETION APPROACH

In this section we estimate the exterior attitude of the cameras starting from the relative attitude measurements. We develop a gradient descent algorithm to minimize a suitable cost function, highlighting its connection with *Matrix Completion* theory. We suppose that the pairwise measurements are subject to noise only. The presence of outliers is handled in Section 5.

4.1 An Introduction to Matrix Completion

The *Matrix Completion* is a well studied problem and appears in many areas other than computer vision, such as collaborative filtering and sensor localization. It consists in recovering the missing entries of a low-rank matrix (Candès and Tao, 2010).

More precisely, the goal is to recover a $n_1 \times n_2$ matrix B of rank $r \ll n_1, n_2$. Only a fraction of its entries is available, represented by index pairs (i, j) in a set $\Omega \subset \{1, 2, \ldots, n_1\} \times \{1, 2, \ldots, n_2\}$. If the number of observed entries is large enough, then solving the following optimization problem

$$\begin{aligned} &\min_X \operatorname{rank}(X) \\ &\text{subject to} \quad \mathcal{P}_\Omega(X) = \mathcal{P}_\Omega(B). \end{aligned} \quad (5)$$

will recover the original matrix correctly. Here, $\mathcal{P}_\Omega$ denotes the orthogonal projection onto the subspace of matrices that vanish outside of Ω. However, this problem is also known to be computationally intractable (NP-hard). An efficient heuristic consists in replacing the rank function in (5) with its convex envelope, that is the *nuclear norm* (Fazel, 2002). The nuclear norm of X is the sum of the singular values of X and it is denoted by $\|X\|_*$. The authors of (Candès and Tao, 2010) proved that under suitable assumptions, with high probability, nuclear norm minimization recovers all the entries of B with no error.

A more practical problem is when the observations are corrupted by noise or the matrix to be reconstructed is only approximately low rank. In this case the constraint $\mathcal{P}_\Omega(X) = \mathcal{P}_\Omega(B)$ must be relaxed, resulting in the following problems

$$\begin{aligned} &\min_X \|X\|_* \\ &\text{subject to} \quad \|\mathcal{P}_\Omega(X) - \mathcal{P}_\Omega(B)\|_F \leq \Theta \end{aligned} \quad (6)$$

$$\min_X \frac{1}{2}\|\mathcal{P}_\Omega(X) - \mathcal{P}_\Omega(B)\|_F^2 \quad \text{subject to} \quad \text{rank}(X) \leq r \tag{7}$$

$$\min_X \lambda\|X\|_* + \frac{1}{2}\|\mathcal{P}_\Omega(X) - \mathcal{P}_\Omega(B)\|_F^2 \tag{8}$$

for some $\Theta \geq 0$ and $\lambda \geq 0$. Here $\|\cdot\|_F$ denotes the Frobenius norm. In this work we focus on problem (7) since in our application the rank of the incomplete matrix is known, as will be explained below. Many authors have proposed efficient methods to solve this problem, such as OPTSPACE, a gradient descent algorithm on the Grassmann manifold (Keshavan et al., 2009).

4.2 Proposed Algorithm

Let $R_{ij} \in SO(3)$ denote the relative rotation between coordinate frames indexed by j and i, and let $\widehat{R}_{ij}$ be an estimate of R_{ij}, obtained through the essential matrix factorization. Only some $\widehat{R}_{ij}$ are known and they are represented by index pairs (i, j) in a set $\mathcal{N} \subset \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$. The goal is to find the absolute rotations $R_i \in SO(3)$ of the cameras such that the compatibility constraint

$$R_{ij} = R_i R_j^T \tag{9}$$

is satisfied $\forall (i, j) \in \mathcal{N}$. In the presence of noise, the pairwise rotations will in general not be compatible. Thus an appropriate minimization problem is

$$\min_{R_1,\ldots,R_n \in SO(3)} \sum_{(i,j)\in\mathcal{N}} \|\widehat{R}_{ij} - R_i R_j^T\|_F^2. \tag{10}$$

Problem (10) is known as *Multiple Rotation Averaging* in computer vision literature (Hartley et al., 2013). Note that the solution is determined up to a global rotation, corresponding to a change in orientation of the external coordinate system. This fact is inherent to the problem and cannot be resolved without external measurements.

In order to rewrite Problem (10) in terms of matrix completion, we introduce the following notations. Let G and R respectively denote the $3n \times 3n$ block-matrix containing the pairwise rotations and the $3n \times 3$ block-matrix containing the absolute rotations, namely

$$G = \begin{bmatrix} I & R_{12} & \ldots & R_{1n} \\ R_{21} & I & \ldots & R_{2n} \\ \ldots & & & \ldots \\ R_{n1} & R_{n2} & \ldots & I \end{bmatrix}, \quad R = \begin{bmatrix} R_1 \\ R_2 \\ \ldots \\ R_n \end{bmatrix} \tag{11}$$

where I denotes the 3×3 identity matrix. It is shown in (Arie-Nachimson et al., 2012) that G can be decomposed as $G = RR^T$, thus it is symmetric, positive semidefinite and of rank 3. Similarly to G we define $\widehat{G}$ as the $3n \times 3n$ block-matrix containing the *observed* pairwise rotations $\widehat{R}_{ij}$ extracted from the estimated essential matrices; it contains zero blocks in correspondence of the missing pairs. The presence of missing data is very common in real scenarios, for example because of occlusions or matching failure. In particular, if two cameras see the scene from different points of view, then there are no corresponding points between them; thus the essential matrix, and hence the relative motion of the pair, can not be computed. Let Ω be the sampling set of $\widehat{G}$.

The rotation estimation problem (10) is equivalent to

$$\min_G \frac{1}{2}\|\mathcal{P}_\Omega(\widehat{G}) - \mathcal{P}_\Omega(G)\|_F^2 \tag{12}$$

where the unknown matrix G should be of the form (11), thus it is required to be symmetric positive semidefinite and to have rank 3. In addition, G should be composed by rotation matrices with identity blocks along its diagonal.

In order to obtain a matrix completion problem (7) we relax these constraints and consider only the rank-3 requirement, obtaining

$$\min_G \frac{1}{2}\|\mathcal{P}_\Omega(\widehat{G}) - \mathcal{P}_\Omega(G)\|_F^2 \quad \text{subject to} \quad \text{rank}(G) \leq 3. \tag{13}$$

This problem can be solved efficiently using the OPTSPACE algorithm. This method decomposes the unknown matrix as

$$G = USV^T \tag{14}$$

where U, S, V have the same dimensions of the factors in the SVD of G. Such a decomposition guarantees that G satisfies $\text{rank}(G) \leq 3$. The minimum with respect to S is easy to calculate, since the cost function is quadratic in S. The minimum with respect to U, V is found by the gradient descent algorithm, using the rank-3 projection of the data matrix as initial datum. See (Keshavan et al., 2009) for details.

An alternative approach to guarantee that the rank requirement is satisfied, is to express the unknown matrix as

$$G = RR^T \tag{15}$$

where $R \in \mathbb{R}^{3n\times 3}$. Such a decomposition guarantees that G is symmetric, positive semidefinite and of rank at most 3, yielding a tighter relaxation. It does not guarantee that G is composed of rotations. This results in the following unconstrained optimization problem

$$\min_R \frac{1}{2}\|\mathcal{P}_\Omega(\widehat{G}) - \mathcal{P}_\Omega(RR^T)\|_F^2. \tag{16}$$

Let $\mathcal{F}$ be the cost function in (16). We propose to minimize $\mathcal{F}$ by using the gradient descent method with line search. The gradient of the objective function is

$$\text{grad}_R(\mathcal{F}) = 2\mathcal{P}_\Omega(RR^T - \widehat{G})R. \tag{17}$$

As for the initial guess, initial values for each R_i are easily found by propagating the compatibility constraint $R_i = \widehat{R}_{ij}R_j$ along a spanning tree of the epipolar graph, starting from a given rotation assumed to be the identity (see Section 5 for definition of the epipolar graph). As regards the stopping criterion, the algorithm ends when the quantity $\|\mathcal{P}_\Omega(\widehat{G}) - \mathcal{P}_\Omega(RR^T)\|_F/\sqrt{|\Omega|}$ is below a given threshold, where $|\Omega|$ is the number of nonzero elements of $\widehat{G}$. Note that our algorithm does not guarantee that the optimization variable R is composed of rotations. Indeed, minimizing the cost function directly in the rotation space $SO(3)$ is a difficult task, as explained in (Hartley et al., 2013), thus a suitable relaxation of such constraint is usually the preferred technique. To be as near as possible to such constraint, we propose to alternate each gradient descent step with a projection onto $SO(3)$ of each 3×3 block in R. The nearest rotation (in the Frobenius norm sense) can be found through singular value decomposition, as explained in (Keller, 1975).

5 CYCLES AND CONSISTENCY

In this section we explain how to detect the outliers among the relative rotations. The presence of false two-view geometries, which generate outliers, is a common situation in real scenarios and it is caused by repetitive structures in the scene. Indeed, these structures can lead to two-view geometries supported by a large number of correspondences, but not reflecting the underlying true geometry.

We consider the *epipolar graph* $\mathcal{G} = (V, E)$ induced by the relative rotations $\widehat{R}_{ij}$ estimated from pairs of views. This graph has a vertex (V) for each camera and edges (E) in correspondence of the available pairwise rotations. We think of $\mathcal{G}$ to be undirected, since $\widehat{R}_{ij}$ is given if and only if so is $\widehat{R}_{ji}$, and $\widehat{R}_{ji} = \widehat{R}_{ij}^T$. If the graph is not connected, the largest connected component is considered only (otherwise it is impossible to estimate the rotations). In order to identify the inconsistent edges, we study the composition of rotation matrices along cycles. More precisely, we consider connected cycles in which every vertex has degree two, i.e. *circuits*. If the error in a cycle, measured as the deviation from identity, is less than a fixed threshold ϵ, then the cycle is supposed to contain inlier edges only. Actually, it may not be, since two outlier rotations may "compensate" such that their wrong contributes vanish, but this is very unlikely to happen in practice. If the cycle error is greater than the threshold, then the cycle must contain at least one inconsistent rotation. To detect such outliers, we improve the algorithm described in (Enqvist et al., 2011).

The authors of (Enqvist et al., 2011) consider a maximum-weight spanning tree, where the weights are the numbers of inlier correspondences, and they analyze cycles formed by the remaining edges. A cycle is kept if the cycle error, normalized by the factor $1/\sqrt{l}$, where l is the cycle length, is small enough; otherwise the non-tree edge is removed. This approach is highly dependent on the chosen spanning tree: if this tree contains an actual outlier, then such a rotation will not be removed, and hence the estimated absolute rotations are wrong. Hereafter, we name this method as the EOK-Algorithm.

To overcome this drawback, we propose a novel algorithm based on the notion of cycle basis. Indeed, cycles in a graph form a vector space over the field $\mathbb{Z}_2$ of dimension $n_E - n_V + n_C$ (Kavitha et al., 2009), where n_E is the number of edges, n_V is the number of vertices, and n_C is the number of connected components of the graph. A basis for the vector space can be constructed by "completing a spanning forest", namely by adding non-tree edges to a spanning forest. Computations on the cycle space are easily carried out by representing cycles as vectors in $\mathbb{Z}_2^{n_E}$. Our goal is to construct a spanning tree formed by inlier edges only. Under this assumption, we can successfully detect the outlier pairwise rotations by using the EOK-Algorithm. We think of $\mathcal{G}$ to be unweighted, since a relative rotation may be correct even if it is generated by a low number of point correspondences. Moreover, we throw away all the edges not belonging to any cycle, since they give no useful information about rotational consistency.

Algorithm 1 describes the overview of our method. The key observation is that any linear combination of inlier cycles will always generate an inlier cycle, while linear combinations of outlier cycles may generate inlier cycles. Thus, in order to obtain an inlier spanning tree, we propose to sum the *outlier* cycles. Clearly, it is computationally intractable to analyze *all* possible combinations. What we propose are two reasonable approaches to choose the combinations that guarantee, with high probability, to extract a spanning tree formed by inlier edges only.

- Sum the inconsistent cycles that have an outlier edge in common, in order to eliminate that edge. Indeed, if two cycles have an edge in common, then their sum does not contain that edge. If the edge is actually an outlier, then, with high probability, the sum will be a consistent cycle.
- Sum the inconsistent cycles in order to connect the connected components of the set of inlier edges. Indeed, if a cycle contains an edge that connects two components and an other cycle does not contain that edge, then their sum connects the two components. If this cycle is actually an inlier, then we can connect the components through inlier data.

In both cases, we sum cycles in pairs of two (and not triplets, quadruples, ...) in order to set a limit on the number of linear combinations.

Algorithm 1 fails when the spanning tree T in Step 2 is constituted by outlier edges only. In this case E_C will be equal to the empty set during all the subsequent steps. To overcome this problem, it is sufficient to restart the algorithm with a different initial spanning tree. The advantage of our approach is that the output E_C is guaranteed to be constituted by inlier edges *only*. In particular, note that E_C is initialized to the empty set in Step 1 of Algorithm 1, while in the EOK-Algorithm it is initialized to a maximum-weight spanning tree, that may contain outliers.

Algorithm 1 Outlier Removal among the Relative Rotations

Input: epipolar graph $\mathcal{G} = (V, E)$, relative rotations $\widehat{R}_{ij}$ extracted from the estimated essential matrices, $(i, j) \in E$, angular threshold ϵ

Output: set E_C of consistent pairwise rotations

1. Initialize $E_C = \emptyset$ and $C_{\text{guess}} = \emptyset$, where C_{guess} denotes the set of outlier cycles.
2. Compute a spanning tree T from E. Form a cycle basis by completing T and classify all the cycles of the basis into inliers (E_C) and outliers (C_{guess}). Eliminate from E all the edges not belonging to any cycle.
3. Compute a spanning forest F from E_C. If F is connected, then the algorithm ends by applying the EOK-Algorithm to E_C with F as input. Otherwise, go to Step 4.
4. Apply the EOK-Algorithm to each connected component of E_C in order to increase the support of E_C or to identify some outlier edges.
5. Sum the cycles in C_{guess} that have an outlier edge in common, in order to eliminate that edge. If the support of E_C has changed after this step, then go to Step 6. Otherwise, go to Step 7.
6. Compute a new spanning forest F from E_C. If F is a spanning tree, then the algorithm ends by applying the EOK-Algorithm to E_C with F as input. Otherwise, apply the EOK-Algorithm to each connected component of E_C, as in step 4.
7. Sum the cycles in C_{guess} in order to connect the connected components of E_C.
8. Compute a new spanning forest F from E_C. If F is a spanning tree, then the algorithm ends by applying the EOK-Algorithm to E_C with F as input. Otherwise, consider the largest connected component of E_C and apply the EOK-Algorithm to it.

6 POSITION RECOVERY: A NEW INTERPRETATION

Once camera attitudes $R_1, \ldots, R_n$ are recovered, we estimate translations $\mathbf{t}_1, \ldots, \mathbf{t}_n$ directly from point matches as explained

in (Arie-Nachimson et al., 2012). This method is based on a factorization of the essential matrix that generalizes the classical one, since it involves the exterior parameters only.

Let E_{ij} denote the essential matrix of the pair (i,j), namely $E_{ij} = [\mathbf{t}_{ij}]_\times R_{ij}$, where $R_{ij} \in SO(3)$ and $\mathbf{t}_{ij} \in \mathbb{R}^3$ describe the relative orientation between view j and i. It is shown in (Arie-Nachimson et al., 2012) that E_{ij} can be expressed as

$$E_{ij} = R_i([\mathbf{c}_i]_\times - [\mathbf{c}_j]_\times)R_j^T \tag{18}$$

where $\mathbf{c}_i \in \mathbb{R}^3$ denotes the absolute location (center) of the $i-$th camera. The advantage of this expression is that pairwise information is no longer required. As explained in (Arie-Nachimson et al., 2012), the epipolar constraint defined by (18) leads to a linear equation for every pair of matching points

$$(\mathbf{c}_i - \mathbf{c}_j)^T(R_i^T\mathbf{p}_i^{(m)} \times R_j^T\mathbf{p}_j^{(m)}) = 0 \tag{19}$$

where $\mathbf{p}_i^{(1)}, \dots, \mathbf{p}_i^{(N_{ij})}$ and $\mathbf{p}_j^{(1)}, \dots, \mathbf{p}_j^{(N_{ij})}$ are N_{ij} corresponding points from images i and j respectively, expressed in normalized image coordinates. Hence a sparse homogeneous linear system is obtained, that can be expressed in matrix form as

$$A\begin{pmatrix}\mathbf{c}_1\\ \dots \\ \mathbf{c}_n\end{pmatrix} = A\mathbf{c} = \mathbf{0} \tag{20}$$

where A is a matrix of dimensions $(\sum_{i,j} N_{ij}) \times (3n)$ whose entries depend on the set of point matches and on the absolute rotations, according to (19). Note that the solution is determined up to a global similarity. Clearly $\mathbf{c}_i = \mathbf{c}_j \ \forall i,j$ (all the cameras share the same position) is a solution to (19). We define *trivial* such a solution. In particular

$$\begin{aligned}\mathbf{c}_i &= (1,0,0)^T \quad \forall i\\ \mathbf{c}_i &= (0,1,0)^T \quad \forall i\\ \mathbf{c}_i &= (0,0,1)^T \quad \forall i\end{aligned}$$

are three trivial solutions of (19), which generate a $3-$dimensional subspace of $\ker(A)$. Thus there exists a *unique* (up to a similarity) *non trivial* solution if and only if

$$\dim(\ker(A)) = 4 \tag{21}$$

and the sought solution is the optimal solution orthogonal to the trivial subspace. Such a solution is given by the eigenvector associated with the fourth smallest eigenvalue of the matrix A^TA, whose dimensions depend on the number of cameras only (not on the number of point matches).

Our contribution to this exterior position recovery method is a theoretical analysis of the linear system (20). We provide a necessary condition for a non-trivial solution to exist, analyzing the epipolar graph $\mathcal{G} = (V,E)$ generated by the images. More precisely, we show that there exists a unique non trivial solution *only if* the epipolar graph is formed by cycles. In other words, in the presence of edges not belonging to any cycle multiple solutions occur.

Consider for simplicity the case $n = 3$ and suppose that the available pairs are $(1,2)$ and $(2,3)$. The epipolar graph is not formed by a simple cycle, since the edge $(3,1)$ is missing. In this case system (20) can be expressed as

$$A\mathbf{c} = \begin{bmatrix}A_{12} & -A_{12} & 0\\ 0 & A_{23} & -A_{23}\end{bmatrix}\begin{pmatrix}\mathbf{c}_1\\ \mathbf{c}_2\\ \mathbf{c}_3\end{pmatrix} = 0 \tag{22}$$

where

$$A_{ij} = \begin{bmatrix}(R_i^T\mathbf{p}_i^{(1)} \times R_j^T\mathbf{p}_j^{(1)})^T\\ (R_i^T\mathbf{p}_i^{(2)} \times R_j^T\mathbf{p}_j^{(2)})^T\\ \dots \\ (R_i^T\mathbf{p}_i^{(m)} \times R_j^T\mathbf{p}_j^{(m)})^T\\ \dots \\ (R_i^T\mathbf{p}_i^{(N_{ij})} \times R_j^T\mathbf{p}_j^{(N_{ij})})^T\end{bmatrix}. \tag{23}$$

We observe that $\text{rank}(A_{ij}) = 2$ or, equivalently, that the $3-$space points $(R_i^T\mathbf{p}_i^{(m)} \times R_j^T\mathbf{p}_j^{(m)})$ lie on a common plane for varying m. Actually, these points lie on a plane which is orthogonal to the baseline of the pair (i,j). To see this, recall that the corresponding points $\mathbf{p}_i^{(m)}$, $\mathbf{p}_j^{(m)}$ and the camera centers lie on a common epipolar plane, and, for varying m, the epipolar planes rotate around the baseline. Consequently

$$\text{rank}(A) = \text{rank}\begin{bmatrix}A_{12} & 0\\ 0 & -A_{23}\end{bmatrix} = 4 \tag{24}$$

since the second block-column of A is a linear combination of the others. Thus $\dim(\ker(A)) = 5$, which means that there exist multiple non trivial solutions to the exterior position recovery problem. These solutions can be computed as follows. Let $\mathbf{b}_{12}$ and $\mathbf{b}_{23}$ respectively denote the baselines of the pairs (1,2) and (2,3), that solve $A_{12}\mathbf{b}_{12} = 0$ and $A_{23}\mathbf{b}_{23} = 0$. By computation we obtain

$$A\mathbf{c} = 0 \Leftrightarrow \begin{cases}\mathbf{c}_1 = \mathbf{c}_2 \ \text{ or } \ \mathbf{c}_1 - \mathbf{c}_2 = \alpha\mathbf{b}_{12}\\ \mathbf{c}_2 = \mathbf{c}_3 \ \text{ or } \ \mathbf{c}_2 - \mathbf{c}_3 = \beta\mathbf{b}_{23}\end{cases} \tag{25}$$

and hence the solutions are

$$\mathbf{c} = \begin{pmatrix}\mathbf{c}_3\\ \mathbf{c}_3\\ \mathbf{c}_3\end{pmatrix},\ \mathbf{c} = \begin{pmatrix}\mathbf{c}_1\\ \mathbf{c}_1\\ \mathbf{c}_1 - \beta\mathbf{b}_{23}\end{pmatrix},\ \mathbf{c} = \begin{pmatrix}\mathbf{c}_2 + \alpha\mathbf{b}_{12}\\ \mathbf{c}_2\\ \mathbf{c}_2\end{pmatrix} \tag{26}$$

for some $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3 \in \mathbb{R}^3$ and $\alpha, \beta \in \mathbb{R}$. The first solution corresponds to the trivial $3-$dimensional subspace of $\ker(A)$. In the second solution cameras 1 and 2 have the same centre, while in the third solution cameras 2 and 3 have the same centre. This is possible since the epipolar graph is not constituted by a cycle, and hence we do not have any compatibility constraint between the baselines.

The discussion above applies equally well to the general case. If there are $n > 3$ cameras and the epipolar graph is not covered by cycles, then $\dim(\ker(A)) > 4$. Indeed, if an edge does not belong to any cycle then the camera centers corresponding to its endpoints can collapse, yielding to multiple solutions. In conclusion, in order to successfully estimate the exterior positions, the epipolar graph is required to be formed by cycles only.

7 EXPERIMENTS

In this section we discuss the efficiency and accuracy of our contributions in both synthetic and real scenarios. All the simulations are carried out in MATLAB on a dual-core 1.3 GHz machine.

7.1 Synthetic Images - Attitude Estimation

We analyze the performances of our matrix completion algorithm in the presence of noise and missing data. Since no outlier is introduced among the relative rotations, Algorithm 1 is not applied here. We compare our method with the techniques described in (Arie-Nachimson et al., 2012), namely *spectral decomposition* (EIG) and *semidefinite programming* (SDP). The former enforces the entire columns of R to be orthonormal, instead of imposing

the orthonormality of each 3×3 block. The latter enforces the matrix G to be symmetric positive semidefinite and to have identity blocks along its diagonal. In our implementation, the MATLAB command *eigs* is used for EIG and the SeDuMi toolbox (Sturm, 1999) for SDP. We also include in our analysis the matrix completion algorithm OPTSPACE, whose MATLAB [1]code has been provided by the authors of (Keshavan et al., 2009). To evaluate the accuracy of attitude estimation, any of the metrics analyzed in (Huynh, 2009) can be used, since they all are bi-invariant and respect the topology of $SO(3)$. In our experiments we choose the *angular* (*chordal*) distance, which takes values in the range $[0, 180^\circ]$.

To generate the ground truth scene and motion we proceed as follows. 200 points with 3D coordinates uniformly distributed in the range $[-5, 5]$ are projected onto $n = 100$ images. The camera locations are sampled at random in the cube $[-30, 30]^3$ far from the point cloud. As for the attitude, the $z-$axes of the cameras point toward the centroid of the point cloud, while the $x-$ and $y-$axes are chosen randomly. Thus each 3D point lies in front of all the cameras, and hence the cheirality constraints are satisfied. For simplicity, we assume that all the cameras have the same calibration matrix ($f = 1000$, $r = 1$, $\gamma = 0$, $p_x = p_y = 500$). Finally, a Gaussian noise with variance between 1 and 10 is added to the image point coordinates. Since no outlier is introduced among the correspondences, RANSAC is not applied here.

We consider a realistic scenario in which a percentage p of the relative rotations is missing. In our experiments we consider the cases $p = 0$, $p = 0.5$ and $p = 0.9$. We also analyze the challenging case in which the number of available $\widehat{R}_{ij}$ is $n - 1$, which is the theoretical *minimum* number of relative rotations necessary to solve for absolute rotations. Figure 1 shows the results, averaged over 30 trials. In the cases $p = 0$, $p = 0.5$ and $p = 0.9$, all the analyzed techniques are equally robust with respect to noise. If the number of available relative rotations is $n - 1$, then EIG and OPTSPACE yield gross errors in the estimates of the attitudes. On the contrary, our method and SDP gives good results even in this challenging case.

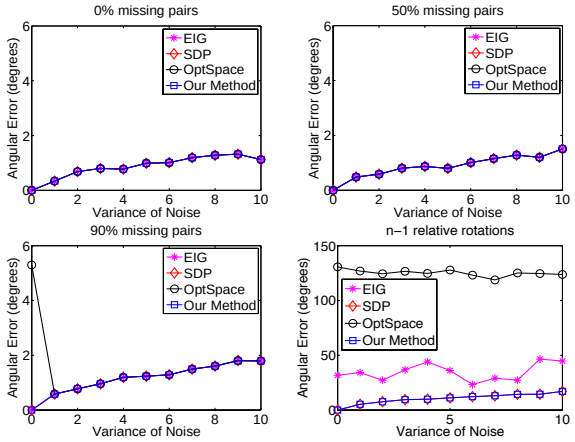

Figure 1: Mean angular errors in estimating the absolute rotations as a function of the variance of Gaussian noise. Note that in the bottom right figure the scale in the $y-$axis is different from the other figures.

We also analyze the efficiency of our method in terms of computational time. Figure 2 reports the running time of the analyzed algorithms as a function of the number of cameras. The execution cost does not include the construction of the data matrix $\widehat{G}$. Our (non-optimized) MATLAB code is significantly faster than semidefinite programming and comparable to EIG and OPTSPACE algorithms.

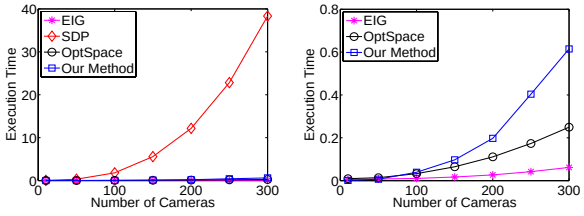

Figure 2: Execution times (in seconds) of exterior attitude estimation as a function of the number of cameras. The right figure is a zoom of the left one.

In summary, our method achieves the best accuracy as SDP but it is much faster.

7.2 Synthetic Images - Outlier Detection

In this section we compare Algorithm 1 with our implementation of the EOK Algorithm. We consider the following figures:

$$\textit{false negative rate} = \frac{c}{a + c} \tag{27}$$

$$\textit{accuracy} = \frac{a + d}{m} \tag{28}$$

where c is the number of *false negatives*, i.e., actual outliers that are erroneously classified as inliers, a is the number of *true positives*, i.e. outliers that are correctly detected, d is the number of *true negatives*, i.e. inliers that are correctly detected, and m is the number of available relative rotations. False negatives are more dangerous than false positives, as they may corrupt the final estimate, whereas false positives can only impact on the statistical efficiency, as they reduce the number of good measurements that are considered in the final estimate. Thus, the main indicator is the false negative rate, which should be as small as possible. The accuracy is considered to check if the method is not rejecting too many samples as outliers.

We consider $n = 20$ images, and we generate the ground truth scene points and camera orientations as done in the previous section. As for the fraction of missing pairs, we analyze the cases $p = 0.25$, $p = 0.5$ and $p = 0.8$. Differently from the previous experiment, the number of matching points is not the same for all pairs, namely each 3D point is not seen by all the cameras. This is done by forcing the visibility matrix to have a band structure. A fraction of the available relative rotations is drawn uniformly from SO(3), simulating outliers. The probability that a given relative rotation is an outlier is inversely proportional to the number of corresponding points of the pair, which is consistent with the assumptions made in (Enqvist et al., 2011).

The angular threshold ϵ is set equal to 3°. Tables 1, 2 and 3 show the results, averaged over 30 trials. Our method yields an effective outlier detection, since false negative rate is always proximal (or even equal) to zero. In the cases $p = 0.25$ and $p = 0.5$, Algorithm 1 gives better results than the EOK Algorithm, as confirmed by the theory. In the case $p = 0.8$, false negative rate of the EOK Algorithm is extremely high, causing misclassification of the outliers, and hence wrong estimates of the absolute rotations. On the contrary, our method yields zero false negative rate, i.e. is an outlier is *never* misclassified as inlier. As for accuracy,

[1] `http://web.engr.illinois.edu/~swoh/software/optspace/code.html`

the EOK Algorithm outperforms our method. The reason why our method is not accurate in this case is that it is not able to find an inlier spanning tree from the *whole* epipolar graph $\mathcal{G}$. It extracts a spanning tree from the largest connected component of $\mathcal{G}$ only, discarding some cameras (and hence some relative rotations).

% outliers	10	20	30	40	50
FNR - our	0	0.003	0.009	0.011	0.015
FNR - EOK	0.057	0.061	0.063	0.056	0.059
AC - our	0.942	0.948	0.937	0.946	0.916
AC - EOK	0.887	0.844	0.797	0.796	0.813

Table 1: Outlier detection: false negative rate (FNR) and accuracy (AC) of the classification in the case $p = 0.25$.

% outliers	10	20	30	40	50
FNR - our	0.022	0.019	0.008	0.016	0.023
FNR - EOK	0.078	0.105	0.108	0.101	0.106
AC - our	0.802	0.782	0.770	0.738	0.693
AC - EOK	0.808	0.757	0.715	0.741	0.749

Table 2: Outlier detection: false negative rate (FNR) and accuracy (AC) of the classification in the case $p = 0.5$.

% outliers	10	20	30	40	50
FNR - our	0	0	0	0	0
FNR - EOK	0.389	0.324	0.318	0.322	0.321
AC - our	0.407	0.445	0.497	0.503	0.509
AC - EOK	0.684	0.701	0.725	0.721	0.726

Table 3: Outlier detection: false negative rate (FNR) and accuracy (AC) of the classification in the case $p = 0.8$.

7.3 Real Images

In this section we apply the techniques presented in this paper to estimate the absolute motion of real cameras, from which the scene structure captured by the images can be recovered (up to a global similarity).

The complete pipeline from the input images, for which the calibration matrices are assumed to be known, to the 3D reconstruction is as follows. First, the SIFT keypoints are extracted and matched to obtain pairs of corresponding points across the n input images. The essential matrices are computed through RANSAC and refined by using Gauss-Newton iterations on the essential manifold, as explained in (Helmke et al., 2007). This method is based on a unique and robust local parameterization of the manifold based on the algebraic properties of the essential matrix. Each essential matrix is factored to obtain a unique pairwise rotation, which is considered missing if insufficiently many inliers are found. Algorithm 1 with $\epsilon = 1^\circ$ is used to detect wrong relative rotations, and all the edges not belonging to any cycle are discarded, because for the corresponding images, the recovery of exterior orientation is not possible with this method. The (inlier) pairwise rotations are used to compute the set of exterior attitudes through the matrix completion algorithm described in section 4. The corresponding points and the recovered rotations are used to solve for the exterior positions through the linear algorithm presented in (Arie-Nachimson et al., 2012). The coordinates of the 3-space points that project to the images are computed by triangulation and image points with high reprojection error are eliminated. Finally, the quality of structure and motion estimation is improved through Bundle Adjustment.

We consider two collections of images for which ground truth calibration and motion are provided, namely the *Fountain-P11* and the *Herz-Jesu-P8* sequences (Strecha et al., 2008). The datasets are formed respectively by $n = 11$ and $n = 8$ images of dimensions 3072×2048 pixels. Results are shown in Table 4 and Figures 3, 4. Our method is able to recover camera positions and orientations accurately, and yields a rich 3D reconstruction of the scenes. The root-mean-square reprojection error is 0.4640 pixels for the Fountain-P11 dataset and 1.0262 pixels for the Herz-Jesu-P8 dataset. The percentages of missing pairs are respectively 25.45% and 32.14%.

	Fountain-P11	HerzJesu-P8
Angular Error - before BA	0.8748°	0.6720°
Angular Error - after BA	0.0516°	0.0607°
Location Error - before BA	$0.1227\ m$	$0.2249\ m$
Location Error - after BA	$0.0037\ m$	$0.0128\ m$

Table 4: Mean errors in estimating the rotations (degrees) and locations (meters) of the cameras, before and after applying bundle adjustment (BA).

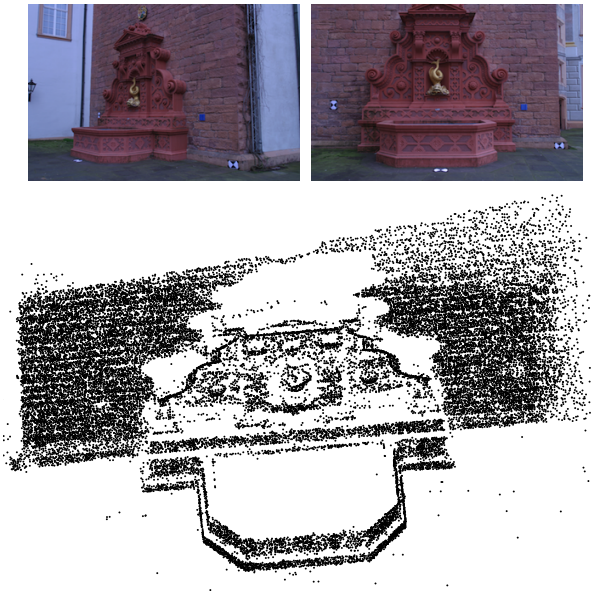

Figure 3: Top: two of 11 images of the Fountain-P11 sequence. Bottom: the sparse 3D reconstruction obtained with our method.

8 CONCLUSION

In this paper we addressed the problem of recovering the attitudes and positions of n cameras in a global SfM system. We proposed a gradient descent algorithm to estimate the attitudes of the cameras, based on low-rank matrix completion, obtaining remarkable results even in the extreme situation where only $n - 1$ pairwise measurements are available. Moreover, we formulated a novel method based on the notion of consistency and cycle basis in order to remove outlier input rotations, improving the technique proposed in (Enqvist et al., 2011). Finally, a theoretical analysis of the exterior position recovery algorithm described in (Arie-Nachimson et al., 2012) was presented.

As regards possible future work, Algorithm 1 could be reformulated using the theoretical formalism of the *Group Feedback Edge Set* problem. As an alternative, we are developing a matrix completion algorithm which is robust both with respect to noise and outliers, avoiding the need of a demanding preliminary outlier rejection step (Arrigoni et al., 2014). Finally, we plan to investigate

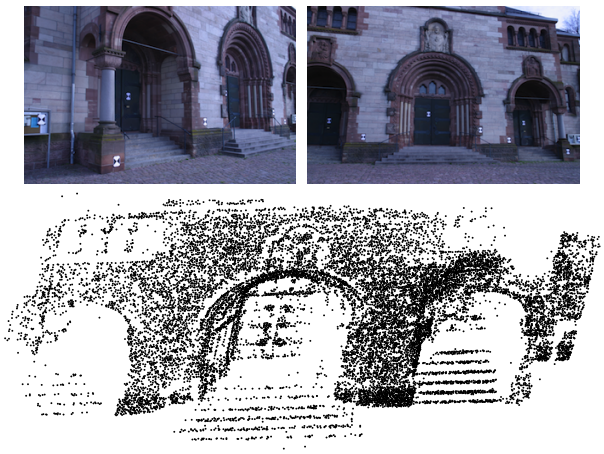

Figure 4: Top: two of 8 images of the Herz-Jesu-P8 sequence. Bottom: the sparse 3D reconstruction obtained with our method.

sufficient conditions under which the position estimation problem admits a unique non-trivial solution.

ACKNOWLEGMENTS

The authors would like to thank Marina Bertolini, Diego Carrera and Luca Magri for their support and useful suggestions. Romeo Rizzi and Carlo Comin pointed out the link between cycle consistency and Group Feedback Edge Set problems. As regards the experiments, we used the MATLAB codes by Andrea Vedaldi (SIFT), Peter Kovesi (RANSAC) and Taehee Lee (Bundle Adjustment).

REFERENCES

Arie-Nachimson, M., Kovalsky, S. Z., Kemelmacher-Shlizerman, I., Singer, A. and Basri, R., 2012. Global motion estimation from point matches. International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission.

Arrigoni, F., Rossi, B., Magri, L., Fragneto, P. and Fusiello, A., 2014. Robust absolute rotation estimation via low-rank and sparse matrix decomposition. Preprint. http://www.diegm.uniud.it/fusiello/papers/rgodec14.pdf.

Brown, M. and Lowe, D. G., 2005. Unsupervised 3D object recognition and reconstruction in unordered datasets. In: Proceedings of the International Conference on 3D Digital Imaging and Modeling.

Candès, E. J. and Tao, T., 2010. The power of convex relaxation: near-optimal matrix completion. IEEE Transactions on Information Theory 56(5), pp. 2053–2080.

Crosilla, F. and Beinat, A., 2002. Use of generalised procrustes analysis for the photogrammetric block adjustment by independent models. ISPRS Journal of Photogrammetry & Remote Sensing 56(3), pp. 195–209.

Enqvist, O., Kahl, F. and Olsson, C., 2011. Non-sequential structure from motion. In: Eleventh Workshop on Omnidirectional Vision, Camera Networks and Non-classical Camera.

Fazel, M., 2002. Matrix Rank Minimization with Applications. PhD thesis, Stanford University.

Gherardi, R., Farenzena, M. and Fusiello, A., 2010. Improving the efficiency of hierarchical structure-and-motion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, pp. 1594 – 1600.

Govindu, V. M., 2001. Combining two-view constraints for motion estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Hartley, R. I. and Zisserman, A., 2004. Multiple View Geometry in Computer Vision. Second edn, Cambridge University Press.

Hartley, R. I., Trumpf, J., Dai, Y. and Li, H., 2013. Rotation averaging. International Journal of Computer Vision.

Helmke, U., Huper, K., Lee, P. Y. and Moore, J., 2007. Essential matrix estimation using Gauss-Newton iterations on a manifold. International Journal of Computer Vision.

Huynh, D. Q., 2009. Metrics for 3D rotations: Comparison and analysis. Journal of Mathematical Imaging and Vision 35(2), pp. 155–164.

Kahl, F. and Hartley, R. I., 2008. Multiple-view geometry under the l^∞-norm. IEEE Trans. Pattern Anal. Mach. Intell. 30(9), pp. 1603–1617.

Kavitha, T., Liebchen, C., Mehlhorn, K., Michail, D., Rizzi, R., Ueckerdt, T. and Zweig, K., 2009. Cycle bases in graphs: Characterization, algorithms, complexity, and applications. Computer Scienze Review 3(4), pp. 199–243.

Keller, J., 1975. Closest unitary, orthogonal and Hermitian operators to a given operator. Mathematics Magazine 48, pp. 192–197.

Keshavan, R. H., Montanari, A. and Oh, S., 2009. Matrix completion from a few entries. ISIT.

Keshavan, R. H., Montanari, A. and Oh, S., n.d. OptSpace: a matrix completion algorithm. `http://web.engr.illinois.edu/~swoh/software/optspace/code.html`.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), pp. 91–110.

Martinec, D. and Pajdla, T., 2007. Robust rotation and translation estimation in multiview reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Moulon, P., Monasse, P. and Marlet, R., 2013. Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion. In: Proceedings of the International Conference on Computer Vision, Sydney, Australie, p. to appear.

Ni, K. and Dellaert, F., 2012. Hypersfm. 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission 0, pp. 144–151.

Snavely, N., Seitz, S. M. and Szeliski, R., 2006. Photo tourism: Exploring photo collections in 3D. ACM Transactions on Graphics 25(3), pp. 835–846.

Strecha, C., von Hansen, W., Gool, L. J. V., Fua, P. and Thoennessen, U., 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Sturm, J. F., 1999. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. Optimization Methods and Software 11-12, pp. 525–653.

Triggs, B., McLauchlan, P. F., Hartley, R. I. and Fitzgibbon, A. W., 2000. Bundle adjustment - a modern synthesis. In: Proceedings of the International Workshop on Vision Algorithms, Springer-Verlag, pp. 298–372.